LLNL-TM-409992

LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

File System-Aware Job Scheduling with Moab

D.A. Lipari, P.D. Eckert

July 8, 2009

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

File System-Aware Job Scheduling with Moab

Many jobs that run on LC systems utilize a parallel file system such as Lustre or GPFS. From time to time, whether for scheduled or unscheduled reasons, these systems are taken offline. In an effort to withhold scheduling jobs that rely on a parallel file system, LC system administrators routinely disable batch scheduling when a parallel file system goes down.

During times of parallel file system maintenance, users who submit jobs that do not require access to parallel file systems have traditionally called the LC Hotline to request that their jobs be given the privilege to run. Hotline staff would then invoke an administrative override to make them eligible for scheduling.

The Moab workload manager now offers the ability to automate this process. Users with jobs that do not require a parallel file system can specify such in their job submission. When a parallel file system goes down, Moab will only schedule those jobs that do not require access to *that* parallel file system.

Here's how it works:

For Users

1. Users who submit jobs that require a parallel file system do not have to do anything different. Moab will assume that jobs with no file system specification require all file systems. If any file system connected to a cluster goes down, Moab will withhold scheduling these jobs.

   Note: Moab's default behavior has changed. Whereas Moab used to blindly schedule jobs that did not specify a file system dependency (and those jobs would subsequently hang when they attempted to access a downed file system), Moab's new handling of jobs that do not specify a file system dependency will be to withhold scheduling those jobs when any file system goes down.

2. Users who submit jobs that require a specific file system (e.g., lscratcha) should specify the dependence on that file system with the following options:

   ```
   msub -l gres=lscratcha <job.command.script>
   ```

   Moab will withhold scheduling such jobs when the specified file system is taken offline. The `gres` designation stands for "generic resource."

3. Users who submit jobs that require multiple file systems (e.g., lscratcha and lscratchb) should specify the dependence on these file systems with the following options:

   ```
   msub -l gres=lscratcha,gres=lscratchb <job.cmd>
   ```

   Moab will withhold scheduling such jobs when any one of these file systems is taken offline.

4. Users who submit jobs that do not require access to any parallel file system would make that declaration with the following msub option:

   ```
   msub -l gres=ignore <job.command.script>
   ```

   Moab will continue to keep these jobs eligible for scheduling no matter what the state of the parallel file systems.

5. When a file system goes down, Moab will automatically hold those jobs that did not specify any file system dependencies (item 1 above). Users who submitted jobs that did not specify any file system dependencies can always modify their jobs' specification to allow Moab to make them eligible for scheduling. They would do so with the following option to the msub command:

   ```
   mjobctl -m gres=ignore <jobID>
   ```

The accepted `gres` values for all LC Linux systems (i.e., not for uP or purple) are lscratch[a,b,c,d] on the OCF and lscratch[1,3] on the SCF.

For System Administrators

1. When a file system is brought down for maintenance, the system administrator runs the following command:

   ```
   mrsvctl -c -R gres=lscratcha -t ALL
   ```

   This has the effect of creating a reservation for lscratcha across the entire grid.

   Note: The omoab grid contains the uP AIX machine which does not mount Lustre. So, the above command needs to be replaced by the following set of commands to target individual machines that mount Lustre if scheduling for uP is to be unaffected:

   ```
   mrsvctl -c -R gres=lscratcha -p hera -t ALL
   mrsvctl -c -R gres=lscratcha -p oslic -t ALL
   mrsvctl -c -R gres=lscratcha -p yana -t ALL
   mrsvctl -c -R gres=lscratcha -p zeus -t ALL
   ```

2. For scheduled downtimes to do file system maintenance, the system administrator adds the start and end times to the reservation creation (for the omoab grid, run the command with the multiple `-p` options as noted in step 1 above):

   ```
   mrsvctl -c -R gres=lscratcha -t ALL
   -s [HH[:MM[:SS]]][_MO[/DD[/YY]]]
   -e [HH[:MM[:SS]]][_MO[/DD[/YY]]]
   ```

3. If required, the administrator cancels all running jobs that are currently trying to access the downed file system(s).

   ```
   canceljob jobID1 jobID2 jobID47
   ```

4. When the file system is brought back online, the administrator releases the reservation(s) created in steps 1 or 2:

```
mrsvctl -r <reservationID>
```

For the LC Hotline Staff

The LC Hotline staff no longer needs to take action to make jobs run when a file system goes down. They will be available to answer questions from users on how users can make their jobs eligible to run.